Comments and Controversies

Why voxel-based morphometric analysis should be used with great caution when characterizing group differences

Christos Davatzikos*

Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104, USA

Received 12 February 2004; revised 1 May 2004; accepted 7 May 2004

A variety of voxel-based morphometric analysis methods have been adopted by the neuroimaging community in the recent years. In this commentary we describe why voxel-based statistics, which are commonly used to construct statistical parametric maps, are very limited in characterizing morphological differences between groups, and why the effectiveness of voxel-based statistics is significantly biased toward group differences that are highly localized in space and of linear nature, whereas it is significantly reduced in cases with group differences of similar or even higher magnitude, when these differences are spatially complex and subtle. The complex and often subtle and nonlinear ways in which various factors, such as age, sex, genotype and disease, can affect brain morphology, suggest that alternative, unbiased methods based on statistical learning theory might be able to better quantify brain changes that are due to a variety of factors, especially when relationships between brain networks, rather than individual structures, and disease are examined.

© 2004 Elsevier Inc. All rights reserved.

Keywords: Voxel; Brain; Morphometric analysis

Computational neuroanatomy is a rapidly evolving field, which utilizes image analysis methods, such as segmentation and deformable registration, to quantify morphological characteristics of different brains from three-dimensional, and often four-dimensional, images. This quantification can subsequently be used for understanding the ways in which age, sex, disease, genetic composition, environmental exposures, treatment, or other factors affect brain structure. It can also be used for diagnostic purposes. Voxel-wise analysis of various morphometric measurements (Ashburner and Friston, 2000; Chung et al., 2001; Davatzikos et al., 1996, 2001; Thompson et al., 1997) has gained attention in the neuroimaging literature (Good et al., 2002; Hulshoff Pol et al., 2001; Job et al., 2002; Kubicki et al., 2002; Mehta et al., 2003), mostly because it does not require an a priori knowledge of how brain structure is affected by disease or other factors, but it determines structural anomalies or differences directly from the data in an unbiased way. Since whole-brain analysis with fine spatial detail based on manual outlining is prohibitively costly for all practical purposes, voxel-based morphometric analysis (VBMA) has been enthusiastically embraced by the neuroimaging community, although it has also been criticized (Bookstein, 2001). In this commentary, to make our presentation concise, we focus on the issue of examining morphological differences between two groups (e.g., a normal control and an affected population). However, our argument has greater generality, including investigations of functional differences between two groups. We argue that VBMA is equipped to answer a question of limited scope, namely “Does a morphological measurement at voxel X differ significantly between two groups?”. Using such voxel-wise statistical tests to make further inferences about how two groups differ, and how such potential differences could be used for understanding the effects of disease on brain morphology as well as for diagnostic or other related purposes, would be a major and poorly founded step, as explained below.

(1) Orientation of group differences in a high-dimensional space. The collection of the voxel-wise morphological measurements constitutes a person’s morphological profile. One can place these measurements into a high-dimensional space, each dimension representing a voxel. If these measurements are first smoothed, then roughly each dimension represents a weighted average of a collection of measurements from the vicinity of a voxel. Group differences are then reflected by the degree of separation of the respective morphological profiles, as Fig. 1 shows schematically. For example, disease or some other factor might shift the measurements obtained from a normal control group (shaded ellipse) by some distance in this high-dimensional space. (In practice, the dimensionality of the space is much higher and equal to the number of voxels being interrogated, but here we use 2D examples for display purposes). Clearly, the larger this shift distance is, the graver the effect of this disease is. However, if the situation is analogous to the one shown on the left of Fig. 1, then VBMA might fail to find this group difference because projection onto each dimension separately results in significant group overlap: diseased and normal groups might have different means, but the overlap is high. If the number of samples is relatively limited, as is often the case, VBMA might detect no group difference along each axis. This is not the case for a situation analogous to the one depicted on the right of Fig. 1, in which the group difference will be detected along Dimension 1 (i.e., Voxel 1), even with a very...
small number of samples. This is a significant bias of voxel-based analysis, which renders inferences about group differences very difficult. In summary, effects that are relatively localized will tend to be detected much easier than effects that are relatively more distributed and involve several structures. This is particularly important when brain networks, rather than structures, are to be examined in relation to a disease.

Of course, if a sufficient number of samples is available, such that the projection of the group differences onto each axis yields significant mean differences, voxel-wise comparisons will still reveal some group differences, as Karl and John correctly point out in their response. However, even if this is the case for some brain regions, it might not be the case for others, depending on the sample size and whether the group difference is strong and focal or more subtle and spatially diffuse. This can introduce a bias in the interpretation of the group differences. More importantly, some brain regions might show higher significance levels than others, which does not necessarily imply that these regions are more important in characterizing underlying group differences.

If they were known in advance, interactions like the ones of Fig. 1 could be sufficiently modeled by treating one region as covariate and examining interactions, as Karl and John indicate in their response. However, this somewhat contradicts one of the principles of voxel-based morphometric analysis, namely the one that does not have a good a priori knowledge of the type of the underlying group differences. Moreover, there are so many ways in which interactions among different brain regions can exist, which renders such an approach impractical, and necessitates a fully multi-variate analysis model.

(2) Distribution overlap as a criterion for group differences. VBMA examines voxel-wise levels of statistical significance, which, as we showed in Fig. 1, are not representative of the overall group separation and difference. Basic principles of statistical learning theory point to alternative measures of significant group differences. For example, a representative of such a measure is the statistical overlap of the two groups, as shown in Fig. 2. Specifically, if one were to draw a Bayesian decision boundary separating the two groups, then a random drawing from the two distributions has probability equal to the darkly shaded region, of being on the wrong side of the decision boundary. This is analogous to the sum of Type I and Type II errors in hypothesis testing, and it reflects the degree by which two groups differ from each other. Statistical learning theory suggests many different ways of forming such decision boundaries, each of which has merits and limitations, depending on the problem at hand, the dimensionality of the measurements, and the available sample (Hastie et al., 2001). For example, in Lao et al. (2003), we adopted a nonlinear support vector machine in conjunction with a hierarchical wavelet-based decomposition of the morphological profiles to examine group differences. We will return to this issue at comment 4 below.

(3) A statistical decision boundary implies a direction in the high-dimensional space; it might, for example, be the direction along which disease or any other factor under study has shifted the morphological profile of a normal population. For simplicity, in Fig. 2 we represent this direction by a single vector, denoted by $w$. Even if one were to define such a vector from voxel-based statistical parametric maps, the possible orientations of this vector $w$ would be severely limited. In the example of Fig. 2, there are only three, out of an infinity of possible orientations, that is, two along the two axes (one of the two voxels shows significant differences) or along the diagonal (both voxels show significant differences). In other words, a continuum of possible morphological differences is reduced to a very limited number of possibilities, each representing a certain number of voxels being “on”, that is, displaying statistically significant differences, and others being “off”.

(4) VBMA typically is based on linear statistics, which in many ways can be very restrictive. For example, distributions in reality often deviate from multi-variate Gaussians, as shown in Fig. 3. In this case, a highly nonlinear hypersurface might be necessary to capture the complex difference between the two groups. Notice that the two vectors, $w_1$ and $w_2$, which summarize the group
differences can be quite different. Most biological structures, including the brain, present highly nonlinear characteristics, thereby implying that nonlinearities analogous to the one of Fig. 3 might be the rule and not the exception. To construct a hypothetical, yet potentially meaningful, example, say that we are interested in understanding the differences between a group that converted from normal to mild cognitive impairment (MCI) and one that did not. Perhaps for people who have a large hippocampus, rate of change might be relatively less important than people who have a small hippocampus to begin with. More generally, the difference between two groups is likely to depend on the morphological profile under consideration and therefore to be inhomogeneous throughout a group.

Diagnosis vs. investigation of spatial maps of morphological group differences

It is often believed that statistical decision methods are only suitable for recognition, as for example in clinical diagnosis, and not for scientific investigation of disease or other processes leading to groups that differ morphologically. Karl and John reiterate this issue in their response. However, this is not necessarily the case (see, for example, Lao et al., 2003). In particular, the vector \( \mathbf{w} \) in Fig. 2 can be used to form a “difference image”: that is, a spatial map of the regions that are most representative of the group differences. This spatial map effectively displays the regions that change when one moves from one side of the dividing hypersurface to the other, that is, from group A to group B in Fig. 2. In the case of nonlinear classifiers, as in Fig. 3, this difference image might vary from one person to another. For example, the group-difference image formed by \( \mathbf{w}_1 \) in Fig. 3 might reflect the anatomical differences between a normal and a diseased population for a particular morphology, and would be different from the group-difference image formed by \( \mathbf{w}_2 \). Although this could be puzzling, at first, it is one of the greatest strengths of nonlinear classification. To further elucidate this issue, we construct a hypothetical example: evaluating the risk of developing dementia might depend not only on the rate of change of the hippocampus and the entorhinal cortex, but also on the size of these structures. For example, it could be that if the hippocampus is relatively small, then the rate of change might be a good predictor of risk of developing dementia, whereas if the hippocampus is large, other morphological characteristics might have higher predictive value.

Focusing on morphological characteristics that matter

One of the strengths of some pattern classification techniques, including support vector machines (Burges, 1998), is that they focus on the interface between two groups, for example, on the boundaries drawn in Figs. 2 and 3, and not on samples that are far away from the dividing boundary (Lao et al., 2003; Golland et al., 2001). This allows them to zoom into the subtleties of group differences, and factor out morphological characteristics that are related to variation within each group and are not useful for distinguishing between two groups. This is not the case for standard voxel-wise statistical tests, such as the ones used by SPM, or by techniques using PCA and related techniques (McIntosh et al., 1996; Strother et al., 1995), and it provides these classification methods with the ability to capture subtle and spatially complex group differences.

Classification vs. regression

In their response, Karl and John make an important point, using as an example the relationship between the parahippocampal gyrus volume and the number of tri-nucleotide repeats in fragile X. The point is that, whenever a relationship between two (continuous in

---

**Fig. 3.** A nonlinear decision boundary might be necessary to characterize the differences between two groups. Notice that the separating vectors \( \mathbf{w}_1 \) and \( \mathbf{w}_2 \) are very different for different morphological profiles. In our previous work, we have found that nonlinear boundaries (Lao et al., 2003) are much more accurate than linear ones in quantifying group differences, implying that the difference between two groups is likely to depend on the morphological profile under consideration and therefore to be inhomogeneous throughout a group.

---

**Fig. 4.** A general framework for analyzing group differences: constructing a diagnostic tool and displaying group differences as spatial maps of regions that best reflect the group differences.
this case) variables is sought, recognition models are not appropriate. Of course, there are regression implementations of support vector machines, which are used for describing such relationships (Hastie et al., 2001). The underlying principles are similar, but the formulations are different. In my commentary, I focussed on investigating group differences and not on regression.

In summary, although voxel-based methods of analysis of morphological (and for the same reasons, of functional) group differences are valuable tools in neuroimaging, they are fundamentally limited. Unless the entire multi-dimensional morphological profile is considered as one large entity, projections onto individual voxels must be treated with caution when inferences are made from them about brain systems affected by disease, sex, genotype, and other factors. The general framework we propose herein involves multi-variate pattern classification in conjunction with image analysis steps, as shown in Fig. 4. We included in some of our discussion a statistical decision methodology that we adopted from Lao et al. (2003); however, other related methods of similar nature can also be found in McIntosh et al. (1996), Miller et al. (1997), and Strother et al. (1995).

Acknowledgments

The author would like to thank Drs. Bilge Karacali and Dinggang Shen for helpful discussions, and grant support by NIH-R01AG14971 and NIH-N01-AG-3-2124.

References
